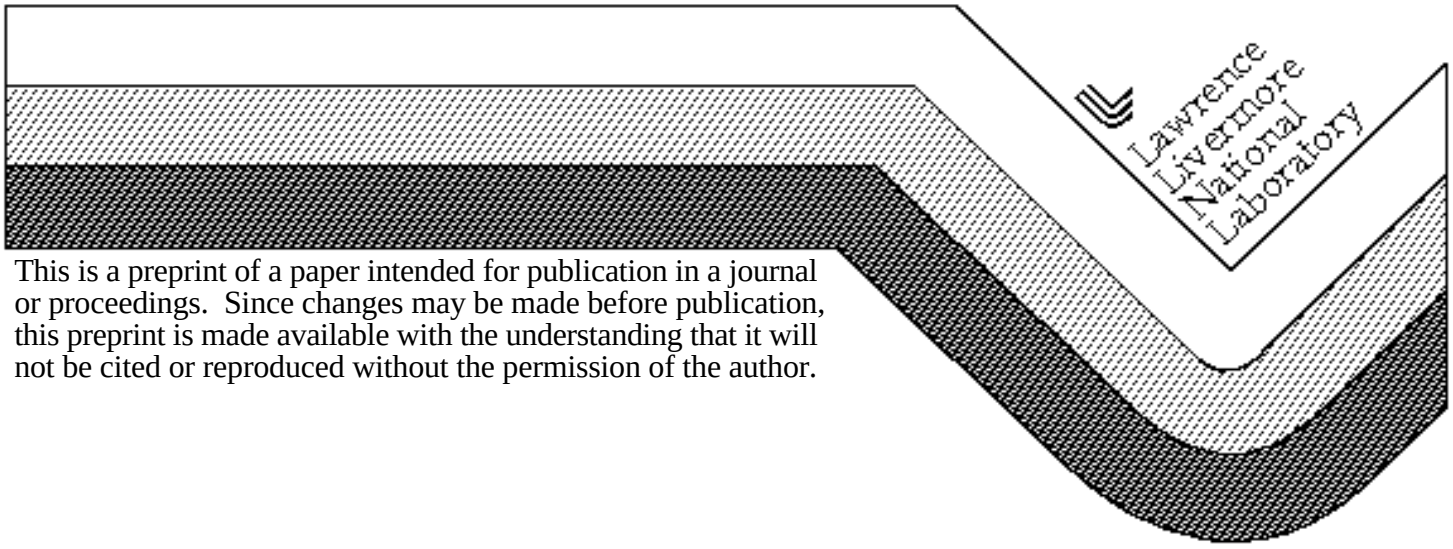
UCRL-JC-123791 Rev 1
PREPRINT

The Key to Enduring Access: Cross-complex Metadata Collaboration

Bruce Lownsbery - Lawrence Livermore National Laboratory,
Helen Newton - Los Alamos National Laboratory,
Axel Ringe - DOE Office of Scientific and Technical Information

This was prepared for submittal to
13th DOE Office Information Technology Conference
Baltimore , Maryland
August 27-30, 1996

August 1996



Work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract W-7405-ENG-48.

The Key to Enduring Access:
Cross-complex Metadata Collaboration

Bruce Lownsbery - Lawrence Livermore National Laboratory,
Helen Newton - Los Alamos National Laboratory,
Axel Ringe - DOE Office of Scientific and Technical Information

13th Office Information Technology Conference, August 27-30, 1996, Baltimore, Maryland

ABSTRACT

The Nuclear Weapons Information Group (NWIG) is a voluntary collaborative effort of government organizations involved in nuclear weapons research, development, production, and testing. Standardized metadata is seen as critical to the locating, accessing, and effective use of the data, information, and knowledge of both past and future weapons activities. This paper will describe the activities of the NWIG Metadata Working Group (MDWG) in developing the metadata elements and authorities which will be used to share information about data stored in computers and vaults across the complex.

With the current lack of secure network connectivity, it is impossible to have distributed access. Therefore we have focused on standardizing the form and content of shared metadata. We have adopted a SGML-based neutral exchange form that is completely independent of how the metadata is created and how it will be used. Our efforts have included the definition of a set of metadata elements that can be applied to all data types and additional attributes specific to each data type, such as documents, drawings, radiographs, photos, movies, etc. We have developed a common subject categorization taxonomy and identified several subsets of a standard glossary and thesaurus for inclusion in the metadata to provide consistency of terminology and the capability to link back to the full thesaurus.

THE SITUATION

Over the course of the past fifty-five plus years, scientists, engineers and a host of technicians and support staff in many locations across the United States have worked diligently to insure our national security through preeminence in nuclear weapons. In order to limit the risk of inadvertent disclosure or access, over the years much of the work has been classified as well as being strategically compartmentalized, with much critical knowledge residing in just a few individual's heads. In the current atmosphere of comprehensive nuclear test bans, reduced budgets, retirements of key individuals, and consolidation of facilities, standardized metadata is seen as critical to locating, accessing, and providing effective use of the knowledge and data of past and future nuclear weapons activities.

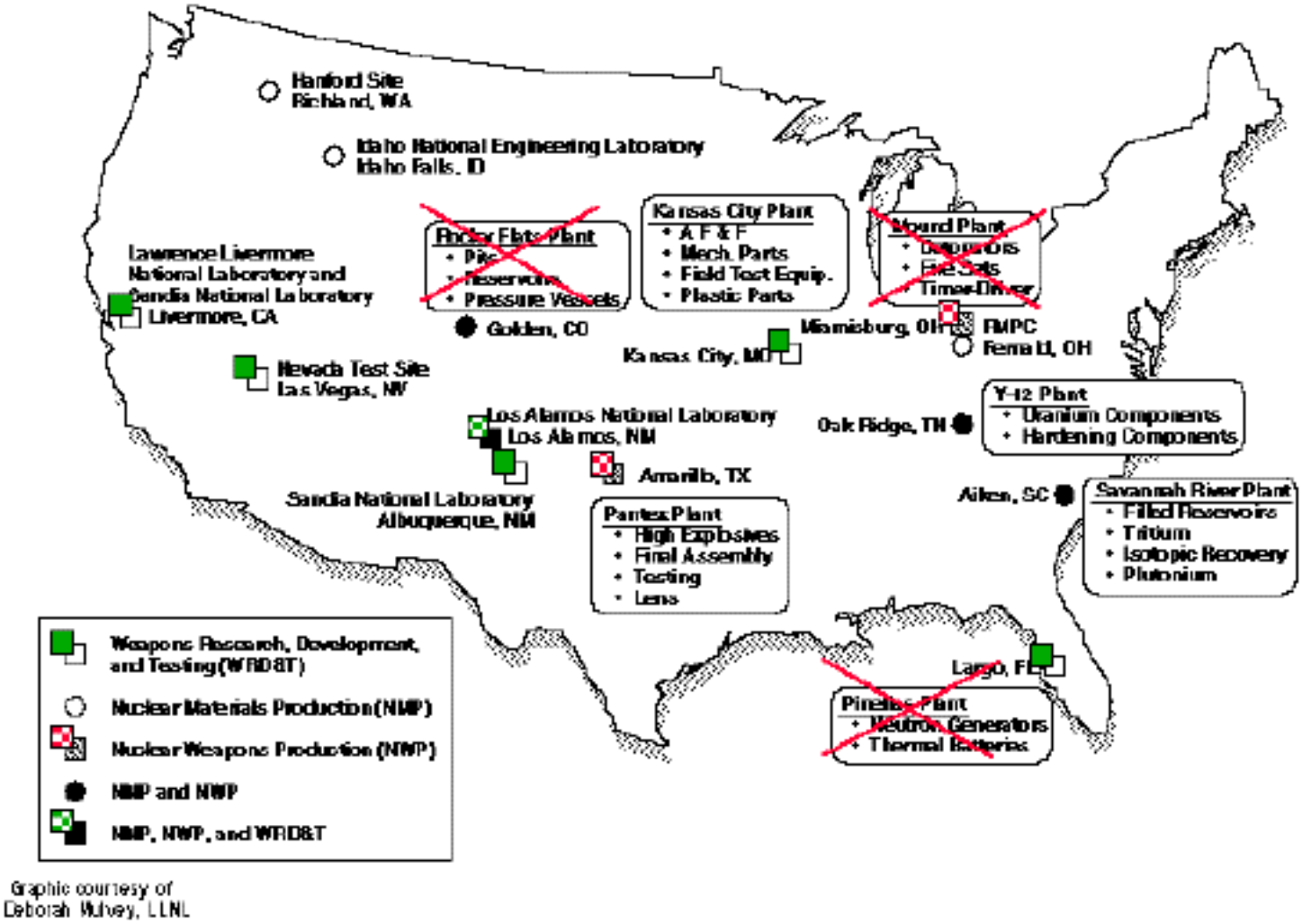
THE NWIG COLLABORATION

The Nuclear Weapons Information Group (NWIG) was conceived late in 1994 to be a voluntary collaborative effort of various government organizations who have an interest in preserving access to information generated during the course of the development of our nuclear deterrent. Participants represent organizations that have been involved in nuclear weapons research, development, production, and testing. At this time we have representatives of various groups from the Department of Energy (DOE), the Lawrence Livermore National Laboratory, the Los Alamos National Laboratory, Sandia National Laboratory, Allied Signal Federal Manufacturing and Technologies, Pantex Plant, Oak Ridge, Bechtel. In addition to the former, all of which are associated with the DOE Weapons Complex, the Defense Special Weapons Agency (DSWA) in the Department of Defense, NASA's Jet Propulsion Laboratory (as an information contractor to DSWA for the DARE Project [1]), and the United Kingdom's Atomic Weapons Establishment are also participants.

Figure 1 provides a map of the US Department of Energy Weapons Complex of the '80s and the general responsibilities of each site, marked to show some of the subsequent changes. The responsibilities and information of the sites that have since been closed have been transferred to others.

WHY METADATA

The NWIG effort is divided into several working groups with each site contributing representatives as they see fit. The current working groups are: Computer Security, Tools, Configuration Control, and Metadata. The NWIG participants realized that the first step to preserving information was to discover what information exists. If all the participants could discover their own information and catalog it in a uniform fashion, it might be possible to share at least the catalogs of the information. The NWIG Metadata Working Group (NWIG-MDWG) was formed and chartered to develop a metadata format for our standardized catalogs of holdings, and is thus the primary contributing group to this paper.



Graphic courtesy of
Deborah Nulvey, LLNL

Figure 1. The US Nuclear Weapons Complex

In the current atmosphere of comprehensive nuclear test bans, reduced budgets, retirements of key individuals, and consolidation of facilities, standardized metadata is seen as critical to the locating, accessing, and effective use of the data, information, and knowledge of both past and future weapons activities. Our metadata is external to the data it represents, rather than being embedded in it. As such, it is potentially less sensitive and, we expect, less constrained by security required compartmentalization which limits access based on what an individual needs to know to do his/her job. Having the metadata external to the file also allows NWIG participants to use the same tools for cataloging the vast amount of the information not yet in an electronic form. Finally, it's relatively small size relative to many of the data objects which are in electronic form, provides for effective search of "catalogs" which are compact enough to be kept on local storage and transferred across networks or by conveniently transportable storage media such as CD-ROM.

METADATA POPULATION AND EXCHANGE

As independent organizations with widely varying situations, the representatives attending the NWIG Metadata Working Group knew we could not rigidly dictate to participants how they should create their catalog information or how they could use the resultant exchanges. Some already had relational databases listing their holdings, others were starting at the beginning with no electronic or paper catalogs --- just shelves or boxes of documents. One of the key elements underlying the efforts of the Metadata Working Group in support of NWIG's goals was to develop a metadata format that would allow the exchange of information by different organizations using different internal cataloging systems. We recognized that each organization would want to have autonomy in creating its own records. Our focus, therefore, was to develop a metadata format that would support exchange of information without requiring organizations to overhaul existing systems (which would usually be impractical) or conform to a common structure that did not meet their institutional needs. By doing this, we would have a standard to ensure that the information exchanged was usable by the recipient. By not dictating to individual organizations the structure of their internal records and databases, we were able to get all participants to agree on a common exchange format. Each compatible database need merely support translation to and from the neutral exchange format. This is much simpler than supporting pairwise translation capabilities for each specific database involved in a given exchange situation, as shown in Figure 2.

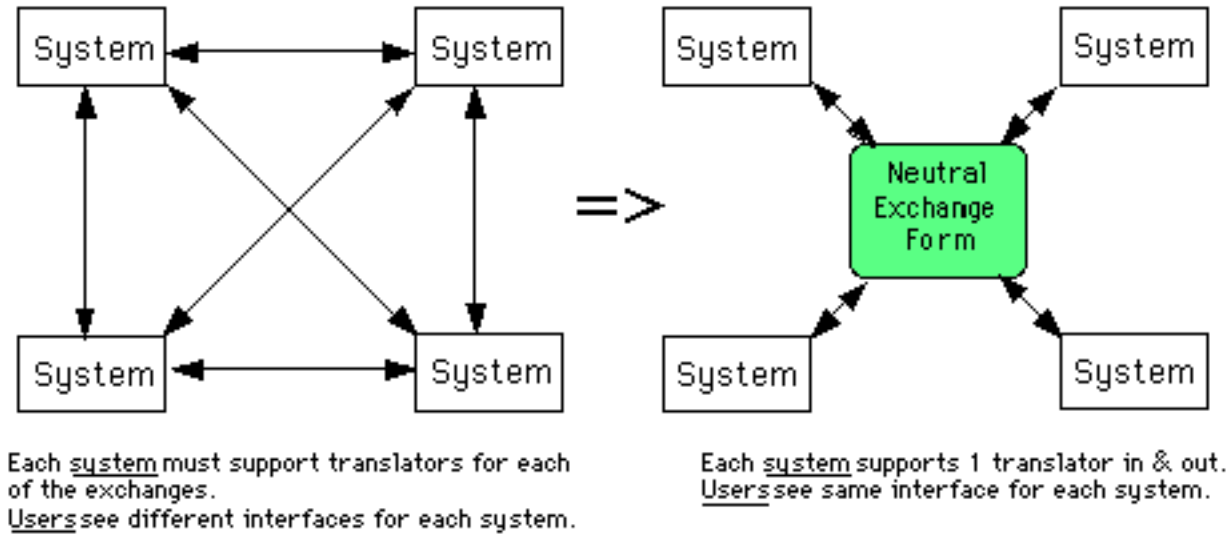


Figure 2. Pairwise Versus Neutral Exchange Format

We emphasize "exchange" because with the current lack of secure network connectivity between sites and the major issue of controlling access, an approach that would allow distributed, rather than direct, access to metadata was set as a goal, but is currently impractical. Our catalogs of metadata will be exchanged on physical media such as CD-ROMs using a format based on Standard Generalized Markup Language (SGML) [2]. SGML provided a ISO standard which supported our requirements for cross-platform computer interpretable metadata with optional, repeatable and variable length entries and the ability to associate attributes with a given entry. It has been adopted by the Department of Energy and the Department of Defense as the standard for conveying document content and has been used in a number of metadata oriented applications as well. The structure and format of a valid exchange file are constrained by the SGML Document Type Definition to ensure that there is a common understanding of what is valid. We have chosen to leave much of the content validation to an external process based on constrained value sets we call "authorities", which will be discussed below. This was to provide additional flexibility in the definition and management of these value sets and their incorporation into other tools.

OVERVIEW OF OUR METADATA

In the process of developing our metadata exchange elements we drew upon the experience of a number of our participants. Our list is not as detailed as that of some other projects, but we believe that our objective is somewhat different. We are proposing to exchange information about data objects which will enable present and future users to understand what is being referred to and to locate information deemed relevant to the ongoing research. Our metadata provides the essence, the details are left to the holder of the information to catalog as they deem necessary. We began the process by surveying a number of existing metadata efforts to look for commonality. While each domain had elements particular to their domain and the function of the associated database, there was a significant base of elements which appeared regularly and these became our starting point.

We have characterized each metadata element as being either mandatory or supplemental. We define mandatory metadata to be that which we will require to be provided. Other desired information, while extremely helpful, may not be readily available or cost prohibitive to generate, and is therefore termed supplemental. A few of the elements are required to be constrained to a single occurrence. Most of the elements are repeatable and we consistently use an SGML attribute (LBL) as a way to label the role that a given instance of a repeated element may play. This saves having to explicitly enumerate all of the possible roles that an occurrence could have, but leaves the labeling of the roles at the discretion of the metadata creator. Such information is considered supplemental and the potential inconsistency that results from creator discretion is therefore outweighed by the simplicity of the approach. In discussing the roles of an "Originating Organization", for example, we could have standardized on roles of 'sponsor', 'prime contractor', etc., but there are a large number of other potential roles that such an organization might play. See, for example, the Originating Organizations element in Appendix A which provides a brief summary of the elements, usage, and SGML samples as they stand at the current draft stage. We have also consistently used an SGML attribute (DEF) as a way to indicate when information was defaulted rather than explicitly entered. This allowed us to include but flag, for example, Creation Dates which may not be known with any certainty. Table 1 below provides the element names, with those which are mandatory distinguished from those which are supplemental. The last element, External Elements is essentially a bucket for embedding fielded elements which are not part of the NWIG standard but which are desired to be preserved as fielded data rather than just Comments.

Mandatory Elements	**Supplemental Elements**
NWIG Identifier	Metadata Sources
Metadata Load Date	Metadata Revisions
Metadata Classification	Object Subtypes
Titles	Media/Format
Creation Dates	Size
Data Object Classification	Attributes
Need-To-Know	Technical Contact
Originating Organizations	External Identifiers
Subject Codes	Authors
Object Types	Keywords
Location	Abstract/ Descriptions
	See Also References
	Browse Forms
	External Elements

TABLE 1. NWIG Metadata Element Names

AUTHORITIES

Providing for enduring interpretable information across many organizations has necessitated the use of common reference points for subjects and terminology. We have developed a common subject categorization taxonomy and identified subsets of a standard glossary and thesaurus for inclusion in the metadata to provide consistency of terminology and the capability to link back to an existing full thesaurus. These authorities are managed by DOE's Office of Scientific and Technical Information as a neutral party. Our initial approach of requiring that keywords be from the controlled thesaurus was deemed an unworkable solution in an environment where we want the end users to be able to create the metadata. A compromise position was to allow uncontrolled keywords and to again use the SGML attribute (LBL) as a way to label that a given keyword conforms to a controlled vocabulary subset. Our current vocabulary subsets cover the names of nuclear explosions (events), nuclear weapons testing series (operations), DOE weapon systems, etc. See, for example, the Keywords element in Appendix A. We have yet to take the additional step of linking local thesauri (which in many organizations are not well documented) to the global thesaurus.

Subject Codes are mandatory elements because it was considered crucial to have a consistent reference frame. This is contrasted with the low-level view of keywords, which are supplemental and not necessarily consistent. In a book analogy, the subject categories correspond to the table of contents and the keywords to the index. Keywords may be gathered from some scan of the text of the title, abstract, content, etc., whereas the subject categories may have to be inferred. Subject categories would be used to locate material which addresses similar topics without requiring that identical terms had been used as keywords, analogous to browsing the stacks around a given Library of Congress number. Because terms within the subject categories are not unique, we discussed several ways to provide the context of the given subject. We have used a system of numerical equivalents to convey the subject taxonomy in a hierarchical structure. Following is a small excerpt from our latest draft:

```
05  NUCLEAR DEVICES AND WEAPONS
         (Includes mockups, scale models, prototypes,
     test devices, and stockpile weapons)
     0501  Gun-Type Devices
     0502  Implosion Devices
          050201  Fission Devices
          050202  Fusion Devices
```

TYPE-SPECIFIC ELEMENTS

We identified many types of data objects to characterize, e.g., documents, drawings, radiographs, photos, movies, video tapes, electronic data in both raw and processed form, output from computer calculations, physical artifacts, etc. Initially it was felt that we would need to provide a set of different elements for each type of data object being characterized. This required standardization on what the subtypes and characteristics of each would be. After struggling with trying to reach agreement on such a diverse set of characteristics, it was observed that this was all supplemental information that would not affect the fundamental purpose of the archive. We have adopted a generic mandatory element, object Type, coupled with recommended values in supplementary elements: Subtype, Size, and Attributes. This provides flexibility to store identifying information without constraining input or requiring conformity. Following is a small excerpt from our latest draft with some of the recommended information:

TYPE	SUBTYPE	SIZE	ATTRIBUTES
Drawing	Part Sketch Map Assembly ...	# sheets (A|B|C|D|E|J)	scale
Photograph	Negative Slide Film-strip ...	physical size	(b/w|color) quality

TOOLS AND COMPUTER SECURITY WORKING GROUPS

Separate NWIG Working Groups are tasked with addressing computer security and tools for both metadata and data. The Tools Working Group collaboration is primarily driven by the need to leverage resources. As well as developing standards for data formats, they are a focal point for sharing methodologies and development efforts. With as diverse a community as NWIG, it is not expected that everyone will use the same tools - the goal is merely that they use compatible tools. The Computer Security Working Group is also driven by the need to leverage resources, but is perhaps even more driven by the need to establish compatible computer security infrastructure at participating sites. They are at the stage of formulating scenarios and requirements for exchanges of metadata and data and are working with the infrastructure providers for implementation of secure networking and associated services.

MDWG SUCCESS FACTORS

Some of the factors that have contributed to the success of the metadata definition effort are:
* Site representatives with a broad cross-section of training, expertise, and background in the weapons environment so that many perspectives were covered and a broad skill base was leveraged.
* An interdependence for data and information and programmatic needs to exchange them so that each site saw value in participating.
* Resource constraints preventing everyone from inventing their own wheel and promoting an environment to leverage other's efforts.
* An approach that currently affects only exchanges and thus allows independence in how organizations create, manage, and utilize metadata internally.
* The information revolution of WWW technology and associated user-driven search which provided a commonly understood vision.
* Free flow of information within the working group through quarterly meetings, WWW pages, e-mail reflectors, etc. so everyone was equally included and informed.
* Overlap in membership of Metadata Working Group with other working groups and main NWIG.
* Meeting locations rotated to various member's sites, promoting understanding of diverse information holdings and origins.
* Participants with cooperative attitudes and a commitment to success so that personalities and pet peeves didn't sidetrack progress.

CONCLUSIONS

Standardized metadata is seen as critical to location, access, and effective use of information and knowledge of both past and future weapons activities. Our multi-organizational effort has initially adopted a metadata exchange approach based on SGML. We have converged on a small number of elements to provide the essence of the data. We have integrated in a set of standard authorities to provide a common reference frame. We have had good cooperation between organizations and have found the collaboration to be quite enjoyable!

ACKNOWLEDGMENTS

We gratefully acknowledge the contributions of individuals from all the NWIG participating organizations for the cooperative development of the NWIG metadata definition and for contributions on this paper.

REFERENCES

[1] R. Borgen, **Data Archival and Retrieval Enhancement (DARE) Metadata Modeling and Representation**, First IEEE Metadata Conference, April 16-18, 1996, NOAA Auditorium, Silver Spring, Maryland

[2] International Organization for Standardization, **ISO 8879: Information processing--Text and office systems---Standard Generalized Markup Language (SGML)**, ([Geneva]: ISO, 1986).

METADATA ELEMENTS
Nuclear Weapons Information Group - Metadata Working Group
DRAFT #5 - 7/14/96

Mandatory elements shown emphasized. Elements proposed to be unclassified are shown with (U).
Repetitions preferred with semicolon delimiters.

Element/ SGML TAG/ Ocurrance	**Definition** (and recommended style/usage information)	**Examples (unrelated to each other)** (distinct examples separated by a line, repeated elements shown concatenated)
NWIG Identifier <M-ID> (U) *ONE ONLY*	Identifier (unique across the NWIG effort) for each data object described in the shared catalog.	<M-ID>LA-01234567</M-ID> <M-ID>LL-AAA79-100000-OO</M-ID>
Metadata Load Date <M-LOAD> (U) *ONE ONLY*	Date that the metadata record was initially loaded into the NWIG catalog. Style: YYYYMMDD	<M-LOAD>19950321</M-LOAD>
Metadata Classification <M-LEV> (U) *ONE ONLY* <M-CAT> <M-WDC> <M-CAV> *ZERO OR MORE*	Four elements which define the classification level, category, sigmas, and caveats of the metadata record. Style: Level and category are from defined list. Sigma categories are listed as semicolon separated numbers preceded by word "SIGMA". NWI categories are written out. All applicable caveats abbreviations are listed.	<M-LEV>U </M-LEV> <M-CAT>RD </M-CAT> <M-WDC>SIGMA1;12</M-WDC> <M-CAV>NOFORN; ORCON</M-CAV>
Metadata Source <M-SRC> *ZERO OR MORE*	The source (author and/or process) which created the metadata about this data object. This information may also be coded in the NWIG-ID number. This element is provided for amplification of that information, if desired. *Freeform	<M-SRC>Vault Custodian </M-SRC>
Metadata Revision <M-REV> (U) *ZERO OR MORE*	Date and notes about revisions to the metadata record. Style: preferred format is date - author - note.	<M-REV>19951102 - Lownsbery, B. - Added author and updated subject codes.</M-REV><M-REV> 19950112 - Lownsbery, B. - Updated keywords.</M-REV>

Title **<TI>** *ONE OR MORE*	The title or subject of the data object. Any security designator is included. Lacking a title security designator (portion marking), assume the classification of the title to be the same as the data object. Note: the titles should be listed in descending order of relevance.	<TI>Diamond Fortune Preshot Report (U)</TI> <TI>H2O-Target assy </TI> <TI>2-Stage Gas Gun Target </TI> <TI>H-Div Equation-of-state.</TI>
Creation Date **<CR-DT>** (𝕌) *ONE OR MORE*	The date of the data object or the date the data object was created. If the date is not fully known (ie: only the year or year and month is known), the element should be completed with zeros and the "def" attribute set to show that it was defaulted. (NOTE: Using the ANSI standard (YYYY-MM-DD) will not permit the entry of zeros where the date is not known.)	<CR-DT>19950321</CR-DT > <CR-DT LBL="formal release">19950321 </CR-DT > <CR-DT def lbl="draft">19930000 </CR-DT >
Data Object Classification **<LEV>** (𝕌) *ONE ONLY* <CAT> <WDC> <CAV> *ZERO OR MORE*	Four elements which define the current classification of the data object. Include any limitation statements. Sigma categories are listed as semicolon separated numbers preceded by word "SIGMA". NWI categories are written out.	<LEV>S</LEV> <CAT>RD</CAT> <WDC>NWI-D</WDC> <CAV>NOFORN; ORCON</CAV>
Need-to-know **<NTK>** (𝕌) *ONE OR MORE*	The "bins" which describe the data content pertinent to need-to-know partitioning. "NULL" is the default value and "NA" indicates Not Applicable, as for unlimited distribution (Caveat DIST-A).	<NTK def>NULL</NTK> <NTK>NA</NTK>
Originating Organization **<ORG>** (𝕌) *ONE OR MORE*	The organization(s) which originated the data object. Repeating this element will allow inclusion of sponsoring and sub-contracting organizations.	<ORG>LANL</ORG> <ORG>LLNL</ORG> <ORG lbl="sponsor">WHITE HOUSE</ORG>
Subject Code **<SUBJ>** *ONE OR MORE*	One or more codes taken from the NWIG subject category list. Mandatory will be at least one entry from the NWIG-approved/OSTI-developed subject codes. Others could be identified by a schema declaration using the "lbl" attribute.	<SUBJ>020043; 056100</SUBJ> <SUBJ>020043; 056100</SUBJ> <SUBJ lbl="orgz">device; materials</SUBJ>
Object Type **<TYPE>** (𝕌) *ONE OR MORE*	The type of the data object, e.g., Document; Video; Audio; Photo; Drawing; Computer Code; etc.	<TYPE>DOCUMENT</TYPE>
Object Subtype <SUBTYPE> (𝕌) *ZERO OR MORE*	A data type specific subtype of the object. For a document object this might be Report, Patent, Speech, etc.	<SUBTYPE>REPORT </SUBTYPE>

Media/Format <MED-FMT> *ZERO OR MORE*	Physical media on which the data object is stored and the data object's format, if applicable. Use a label attribute to convene additional information, such as archive or record copy.	<MED-FMT>PAPER; FILM-35mm </MED-FMT> <MED-FMT lbl="original">FILM </MED-FMT> <MED-FMT lbl="1st generation print">PAPER </MED-FMT>
Size <SIZE> (U) *ZERO OR MORE*	A data type specific measure of size. Ex: for a document = # pages, but for a video = # minutes.	<SIZE>27 p</SIZE> <SIZE>64 MINUTES</SIZE>
Attribute <ATTR> *ZERO OR MORE*	A data type specific set of attributes. Ex: for a video = b/w or color, but for software = language	<ATTR>COLOR </ATTR> <ATTR>FORTRAN </ATTR>
Location **<LOC>** (U) *ONE OR MORE*	The physical and/or electronic locations of the enduring data object. Note: Use of entity definitions are recommended for common locations. The entity definition of the tag should include custodial contact information for the referenced location.	<LOC>X-DO VAULT; DOE-PIT DATA COLLECTION; INSP REC VAULT </LOC> <LOC>http://www.whatever </LOC> <LOC>&LANL-XDO;</LOC>
Technical Contact <CON> (U) *ZERO OR MORE*	Technical contact, distinct from the location contact (see above), for the data object.	<CON>&NWIG-CHR;<CON>
External Identifier <EX-ID> (U) *ZERO OR MORE*	ID by which the originating and other organizations know the data object.	<EX-ID>X-94-003; P-23-94-22</EX-ID> <EX-ID lbl="ERC">AAA79-100000-OO</EX-ID>
Author <AU> (U) *ZERO OR MORE*	The author(s) of the data object. Preferred format is LASTNAME, F.M.	<AU>Richter, J.T.</AU> <AU>Worlton, L. R., etal </AU> <AU>Richter, J.T., Worlton, L. R., etal; Mortensen, F.N.</AU>
Keyword <KW> *ZERO OR MORE*	Non-restricted keywords provided to assist the user in retrieving the data item. Encourage the use of this element to identify one or more nuclear shots and/or systems that relate to this data object.	<KW>DNA; preshot </KW> <KW lbl=EVENT>TAFI </KW>
Abstract/ Description <AB> *ZERO OR MORE*	The abstract or a description of the data object.	<AB>Page 2 of a J size drawing; This report is the official assembly record for test unit XYZ.</AB>

See Also Reference <SEE> *ZERO OR MORE*	Related data objects.	<SEE>P-22-94-303, Rev 1 </SEE> <SEE lbl="preshot report">LL-COMW-90-1234 </SEE>
Version Note <VER> *ZERO OR MORE*	The version of the data object with any notes (possibly the date, and purpose) pertaining to the revision	<VER>Rev B, adding data block, 19930202 </VER>
Comment <COM> *ZERO OR MORE*	Undefined notes about the data object. Character string showing the date of the comment and the individual commenting.	<COM>19950625 - Helen Newton, X-DO Vault Custodian - This object is a one of a kind relic that should be preserved for its historical significance.</COM>
Collection Title <C-TI> *ZERO OR MORE*	Identity of the 'intellectual' collection to which the data object belongs.	<C-TI>Proceedings of XIV Annual DOE HE Conference </C-TI>
Browse Form <BROWSE> *ZERO OR MORE*	URN link to a convenient form for on-line browsing of a representation of the data object.	<BROWSE>http://www.whatever /file.pdf </BROWSE>
External Element <EX-EL> *ZERO OR MORE*	Metadata elements imported from an external system, which do not fit into the standard NWIG elements, but which the metadata author feels deserve preservation and dissemination through NWIG. Note: external elements embedded within this elements are the only ones allowed to have only a start tag an end tag since all other markup is ignored up to </EX-EL>.	<EX-EL><author>748900 2 <accno>526859 <destdate>0287 </EX-EL> <EX-EL><PDS-VERSION-ID >DARE01 <DISTRIBUTION-ID>D <REASON-DATE-TEXT>"Critical Technology, 12 May 1994" <RELEASE-AUTHORITY-NAME>"Defense Nuclear Agency (DNA/DFIM)" <CONTRACT-ID>"DNA 001 75 C 0222" </EX-EL>